Predicting Beijing Housing Price

Cinny Lin (ycl461), Yihan Xu (yx1708), Yizhou Lu (yl5438)

December 7, 2020

1 Introduction

The skyrocketing housing price in metropolis has become a growing concern for people in recent years. Our common sense suggests that house price may be subject to factors such as its location and age. However, information listed on housing websites often times is not enough to explain housing price. Therefore, it is of practical significance to study the factors that affect house prices in big cities, and to what extent the information available online can explain them.

1.1 Research Question

This paper looks into a real-world dataset of 10,682 observations about Beijing's housing prices in 2016, obtained from Lianjia, a primary real estate dealer in China. In the dataset, we noticed that the per square meter housing prices in Beijing is highly inconsistent, with some as reaching as high as 150,000, and others as low as 10,000. Although such inconsistency is not uncommon in other big cities like Shanghai or Shenzhen, housing prices in Beijing are particularly worth investigating, due to two reasons: the coexistence of modern apartment complexes and ancient bungalow houses (Siheyuan), and the concentric city layout. We believe that these characteristics may lead to different sets of variables affecting Beijing's housing prices disproportionately compared to other big cities. Thus, we pose the question: **what are the key factors that influence per unit housing prices in Beijing?**

1.2 Choose Variables

To offer an insight to our proposed research question, we will use the house price per unit ($price$) as the dependent variable.

The independent variables we will use are: latitude (Lat) and longitude (Lng) of the house, size of the house ($square$), number of different rooms in the house ($livingRoom$, $drawingRoom$, $bathroom$, and $kitchen$), dummies of different building types ($buildingType1/2/3/4$), dummies of building types ($buildingType1/2/3/4$) , time of construction ($constructionTime$), dummies of renovation conditions ($renovationCondition1/2/3/4$), , average number of ladders residents share per floor ($ladderRatio$), dummy for elevator availability ($elevator$), dummy for five years property ($fiveYearsProperty$), and dummy of subway proximity ($subway$).

In order to minimize any omitted variable bias, we will first include all possible independent variables in our model, and then iteratively run the model to reduce the model to include only the most significant variables. Meanwhile, we will leave out some irrelevant or ill-specified variables, such as: url of the data, transaction id, community id, trade time (since only 2016 data is included), followers, total price.

Following detailed data anaylsis and comprehensive testing, we finally decided to include 18 independent variables in our model. In the next section, we will do a descriptive analysis and correlation analysis on some of these variables to prepare for our model.

1.3 Literature Review

To construct a reasonable model, we read extensively to understand the methodologies and findings from past researchers. We learned that most researchers use multivariate linear regression model with hedonic pricing model to predict house prices.

We also found that the structural characteristics, such as floor area and number of rooms (Fletcher, et.al., 2000; Li and Brown, 1980) are strongly related to price. Additionally, the age of the house has a negative correlation with the house price as a result of an increase in maintenance cost (Kain and Quigley, 1970).

Location has also been recognized as a main factor for house price. Gelfand et.al. (2004) highlighted the importance of the spatial component in explaining prices. Chen and Hao (2006) further used the distance from houses to the central business district (CBD) as a measurement for the importance of location attributes on housing prices.

In addition to structural and locational factors, neighborhood specific-factors have also been identified as important factors for explaining house prices (Yusof, 2012). Since neighborhood attributes, such as occupation of residents or crime rate, are often affected by socioeconomic factors, it is somewhat more challenging to measure them.

In summary, we can understand housing prices as being determined by three broad categories: structural, locational, and neighborhood factors. However, we must also acknowledge the complexity of measuring the housing price. Therefore, a cautious interpretation of the empirical results is necessary for our project.

2 Variables Description and Relationships

2.1 Descriptive Analysis

2.1.1 Basic Descriptive Statistics

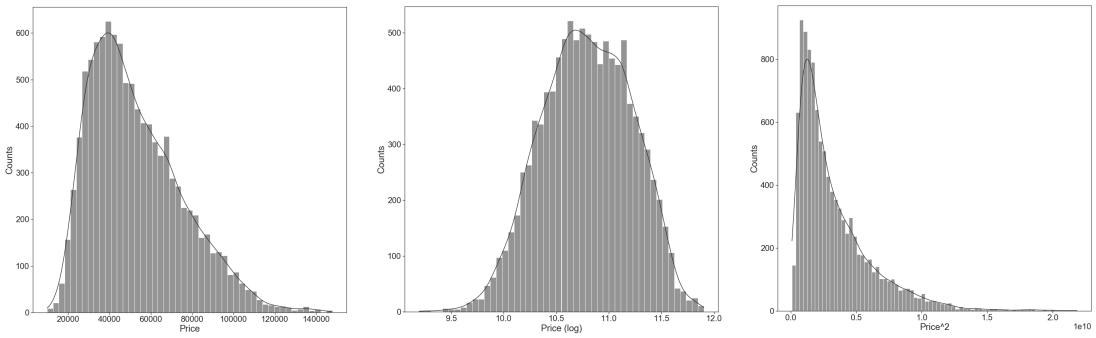
Some basic descriptive statistics of the variables are presented in Table 1.

N = 10331, Beijing, China, 2016

	Min	LQ	Med	UQ	Max	Mean	Std.dev	Skewness
price	9841	35891	48659	67151	147797	53082.72	22340.22	0.81
square	13.7	59	77.73	105.91	458	89.79	46.3	2.17
Lng	116.07	116.34	116.41	116.47	116.71	116.41	0.1	0.07
Lat	39.63	39.9	39.94	40	40.25	39.95	0.095	0.29
livingRoom	0	2	2	3	9	2.1	0.86	0.86
drawingRoom	0	1	1	2	5	1.19	0.57	0.36
kitchen	0	1	1	1	2	0.99	0.12	-3.54
bathRoom	0	1	1	1	6	1.24	0.53	2.35
constructionTime	1944	1994	2002	2006	2016	1999.89	8.97	-0.92
ladderRatio	0.01	0.25	0.33	0.5	2	0.38	0.19	2.07
elevator	0	0	1	1	1	0.61	0.49	-0.43
fiveYearsProperty	0	0	1	1	1	0.62	0.49	-0.49
subway	0	0	1	1	1	0.62	0.48	-0.51
*buildingType*1	0	0	0	1	1	0.28	0.45	0.99
*buildingType*2	0	0	0	0	1	0.0004	0.02	45.42
*buildingType*3	0	0	0	0	1	0.19	0.39	1.58
*buildingType*4	0	0	1	1	1	0.52	0.50	-0.09
*renovationCondition*1	0	0	0	0	1	0.07	0.25	71.85
*renovationCondition*2	0	0	0	0	1	0.03	0.17	5.44
*renovationCondition*3	0	0	0	1	1	0.37	0.48	0.54
*renovationCondition*4	0	0	1	1	1	0.53	0.5	-0.12
*buildingStructure*1	0	0	1	1	1	0.53	0.5	71.85
*buildingStructure*2	0	0	0	1	1	0.33	0.47	0.71
*buildingStructure*3	0	0	0	0	1	0.0004	0.02	45.42
*buildingStructure*4	0	0	0	0	1	0.04	0.21	4.37
*buildingStructure*5	0	0	0	0	1	0.001	0.03	20.29
*buildingStructure*6	0	0	1	1	1	0.62	0.49	-0.5

Table 1: Descriptive Statistics of Variables

2.1.2 Per Unit Price

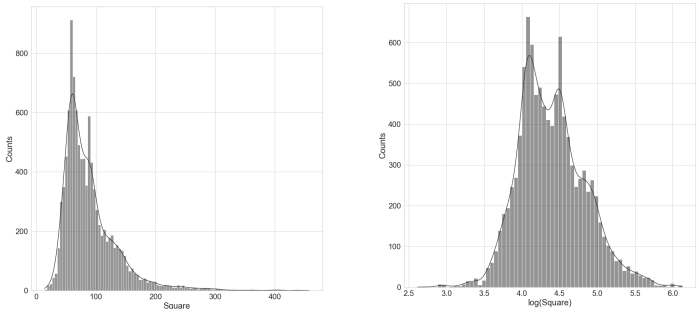


(a) *price* (b) *log*(*price*) (c) $price^2$

Figure 1: Different Interpretations for Price

The three histograms show three distributions of price per square unit for our housing dataset after applying different functions. We decided to use *log*(*price*) for three reasons. First, according to the literature review we did, applying log function to house price is the convention. Second, we observed that applying the log function brings the originally skewed price distribution of closer to normal distribution. Third, using the log form of price is better to interpret how our independent variables influence the percentage change in price.

2.1.3 House Size (Square)

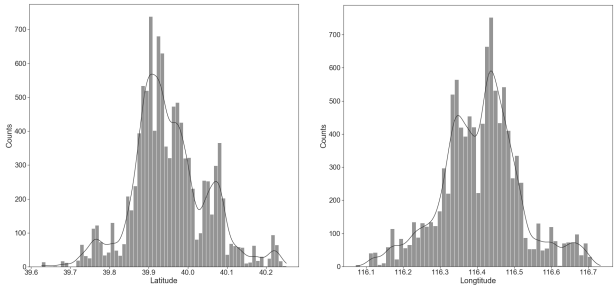


(a) Square distribution (b) log(square) distribution

Figure 2: Distribution of house sizes

The two histograms above show the distribution of house size in square meters, which is in the range of [0,450]. Given our observation of the histograms, we take the log of *square* to adjust the right-skewness. It is also common sense to take the log of house sizes since the interpretation is that an extra $\alpha\%$ in size leads to a $\beta\%$ decrease in per unit price due to high maintenance cost, unsuitability to small- to medium-sized families, etc.

2.1.4 Latitude and Longitude

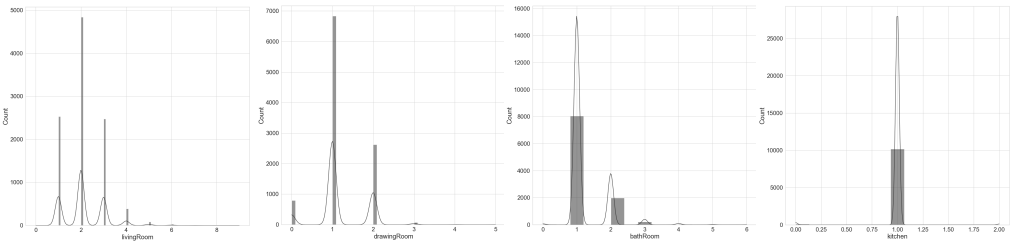


(a) Latitude (b) Longitude

Figure 3: Distribution of latitude and longitude

From the two graphs above, we see that most houses are located in the middle. Since Beijing's city layout is concentric, we can infer that most houses are located close to the city center.

2.1.5 Rooms



(a) livingRoom (b) drawingRoom (c) bathRoom (d) kitchen

Figure 4: Number of rooms in the house

Above shows the distributions of different rooms across houses. We see that the majority of the houses have 2 living rooms, 1 drawing room, 1 bathroom, and 1 kitchen. This makes sense since most houses are near the city and of smaller sizes, they are unlikely to have many rooms.

2.1.6 Construction Time

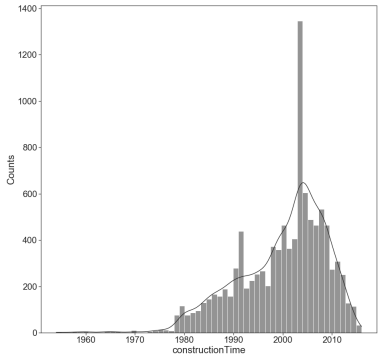
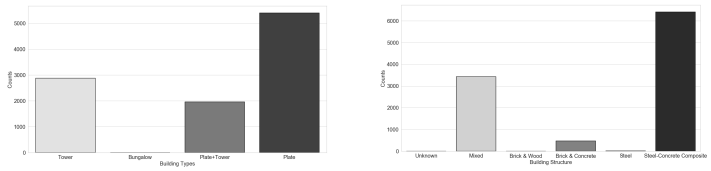


Figure 5: Construction Time counts

Above shows the distribution of construction time, which is in the range of [1944,2016]. We can see that the histogram is left skewed, which means most houses were built relatively recently.

2.1.7 Building Type and Structure



(a) Building Type (b) Building Structure

Figure 6: Distribution of Building Type and Building Structure

In the left graph, we notice most building types are plate and tower, which aligns with the right graph where they are made of mixed and steel-concrete composite materials. We know by common sense that these building types and structures are most likely built near the city center.

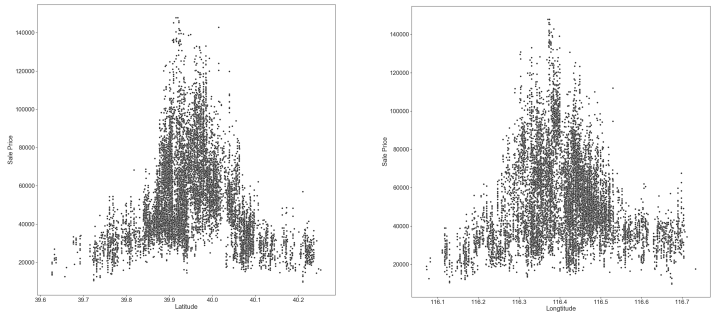
2.2 Correlation Analysis

2.2.1 Correlation Coefficient Matrix

	price_~g	Lng_di~d	Lat_di~d	square~g	living~m	living~2	drawin~m	drawin~2	kitchen	kitchen2	bathRoom	bathRo~2	constr~f	constr~2	ladder~o	elevator	fiveYe~y	subway	buil~pe2	buil~pe3	buil~pe4	renova~2	renova~3	renova~4	buil~re2	buil~re3	buil~re4	buildi~5	buildi~6
price_log	1.0000																												
Lng_diff_q~d	-0.4474	1.0000																											
Lat_diff_q~d	-0.4424	0.1462	1.0000																										
square_log	-0.2158	0.0566	0.1369	1.0000																									
livingRoom	-0.1004	0.0244	0.1043	0.7487	1.0000																								
livingRoom2	-0.1072	0.0173	0.1022	0.7003	0.9503	1.0000																							
drawingRoom	-0.0930	0.0499	0.1061	0.6766	0.5285	0.4914	1.0000																						
drawingRoom2	-0.1178	0.0546	0.1336	0.6524	0.5111	0.5098	0.9354	1.0000																					
kitchen	0.0284	0.0202	-0.0165	0.1453	0.1434	0.1230	0.1789	0.1327	1.0000																				
kitchen2	0.0164	0.0088	0.0041	0.1450	0.1437	0.1407	0.1463	0.1350	0.8683	1.0000																			
bathRoom	-0.1040	0.0110	0.1041	0.7001	0.6088	0.6477	0.5737	0.6178	0.1498	0.1616	1.0000																		
bathRoom2	-0.0893	0.0083	0.0904	0.6442	0.5801	0.6526	0.5171	0.5806	0.1184	0.1534	0.9588	1.0000																	
constructi~f	0.2892	-0.1986	-0.1922	-0.3535	0.0022	-0.0408	-0.2036	-0.2566	0.0349	-0.0044	-0.2597	-0.2299	1.0000																
constructi~2	0.2782	-0.1666	-0.1759	-0.3517	-0.0059	-0.0460	-0.2146	-0.2507	0.0268	-0.0044	-0.2445	-0.2136	0.9534	1.0000															
ladderRatio	-0.0520	0.0690	0.1050	0.4236	0.3417	0.3104	0.3453	0.3137	0.1014	0.1011	0.3220	0.3027	-0.1615	-0.1412	1.0000														
elevator	0.0819	-0.0205	-0.1220	0.1626	-0.0942	-0.0670	0.0298	0.0740	-0.0472	-0.0184	0.0987	0.0789	-0.4916	-0.4328	-0.0808	1.0000													
fiveYearsP~y	0.0458	-0.0772	-0.0694	-0.0154	0.0244	-0.0022	0.0106	-0.0258	0.0345	-0.0020	-0.0535	-0.0583	0.2259	0.1553	-0.0240	-0.1380	1.0000												
subway	0.3234	-0.3058	-0.2040	-0.1503	-0.1206	-0.1104	-0.1388	-0.1381	-0.0304	-0.0235	-0.1098	-0.1007	0.1645	0.1515	-0.1153	0.1139	0.0273	1.0000											
buildingTy~2	0.0417	-0.0121	-0.0106	-0.0562	-0.0194	-0.0149	-0.0322	-0.0185	-0.0819	-0.0461	-0.0373	-0.0152	0.0382	0.0417	-0.0180	-0.0244	-0.0250	0.0153	1.0000										
buildingTy~3	0.0626	-0.0088	-0.0495	0.0190	-0.0928	-0.0809	-0.0077	0.0132	-0.0059	0.0041	-0.0048	-0.0137	-0.2291	-0.1951	0.0447	0.3595	-0.0599	0.0421	-0.0095	1.0000									
buildingTy~4	-0.0810	0.0877	0.1696	-0.0262	0.1460	0.1090	0.0641	0.0181	0.0497	0.0195	0.0125	0.0074	0.2612	0.2589	0.2607	-0.6972	0.0513	-0.1513	-0.0206	-0.5062	1.0000								
renovation~2	-0.0564	0.0784	0.0265	0.0739	0.0844	0.0805	0.0444	0.0445	0.0140	0.0221	0.0643	0.0721	-0.0431	-0.0033	0.0558	0.0059	-0.0668	-0.0541	-0.0035	-0.0145	0.0420	1.0000							
renovation~3	-0.0002	-0.0383	-0.0588	-0.1383	-0.0088	-0.0253	-0.1030	-0.1364	-0.0090	-0.0184	-0.1290	-0.1134	0.2882	0.2429	-0.0686	-0.1943	0.1310	0.0260	-0.0049	-0.0895	0.1046	-0.1363	1.0000						
renovation~4	0.0757	0.0054	-0.0160	0.0979	-0.0160	-0.0033	0.0907	0.1136	0.0243	0.0176	0.0960	0.0761	-0.2191	-0.1982	0.0482	0.1747	-0.0632	0.0121	-0.0110	0.0817	-0.1001	-0.1888	-0.8128	1.0000					
buildingSt~2	-0.0546	-0.0075	0.1232	-0.1572	0.0501	0.0155	-0.0520	-0.1052	0.0403	0.0096	-0.1274	-0.1192	0.4466	0.3762	0.0595	-0.7331	0.1328	-0.0786	-0.0139	-0.2843	0.5540	-0.0171	0.1831	-0.1576	1.0000				
buildingSt~3	0.0417	-0.0121	-0.0106	-0.0562	-0.0194	-0.0149	-0.0322	-0.0185	-0.0819	-0.0461	-0.0373	-0.0152	0.0382	0.0417	-0.0180	-0.0244	-0.0250	0.0153	1.0000	-0.0095	-0.0206	-0.0035	-0.0049	-0.0110	-0.0139	1.0000			
buildingSt~4	-0.0110	-0.0074	0.0253	-0.0394	0.0626	0.0586	-0.0151	-0.0205	0.0114	-0.0018	-0.0054	0.0025	0.1930	0.1900	0.0256	-0.2602	0.0391	-0.0201	-0.0043	-0.1030	0.1868	0.0071	0.0539	-0.0530	-0.1537	-0.0043	1.0000		
buildingSt~5	-0.0215	-0.0053	0.0375	0.0501	0.0711	0.1134	0.0428	0.0708	-0.0222	-0.0135	0.0702	0.1050	0.0024	0.0048	0.0478	-0.0248	-0.0200	-0.0087	-0.0007	-0.0165	-0.0015	0.0104	-0.0202	0.0151	-0.0241	-0.0007	-0.0074	1.0000	
buildingSt~6	0.0581	0.0104	-0.1326	0.1681	-0.0796	-0.0474	0.0547	0.1062	-0.0391	-0.0057	0.1227	0.1080	-0.5178	-0.4485	-0.0716	0.8265	-0.1436	0.0854	-0.0252	0.3209	-0.6165	0.0130	-0.1989	0.1748	-0.9024	-0.0252	-0.2786	-0.0436	1.0000

Figure 7: Correlation Coefficient Matrix

2.2.2 Latitude and Longitude

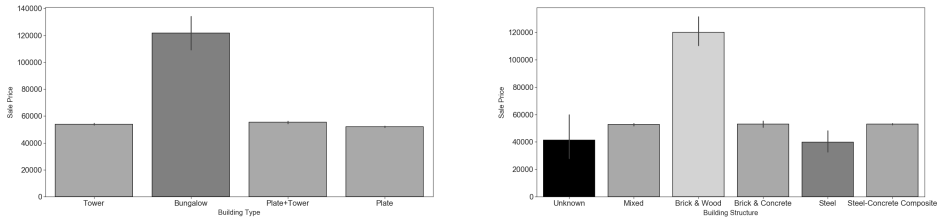


(a) Latitude and Price (b) Longitude and Price

Figure 8: Location (Lat, Lng) and Price

In these two graphs, we can see that there is a wider price range in the middle near the city center, and prices drop significantly as we move away from the city center. Building on these observations, we decided to use $(x - median(x))^2$ as our regressors for the following reasons. Due to Beijing's concentric city planning, by subtracting latitude/longitude by its median we can get its *distance* to the city center. Furthermore, since the correlation between price and distance is closer to a bell shape, we think using $distance^2$ may better explain its distribution.

2.2.3 Building Type and Building Structure



(a) Building Type and Price (b) Building Structure and Price

Figure 9: Building Type and Structure with Price

We observed that the Bungalows on the left graph and the Brick & Wood structures on the right graph both have distinctively different pricing than the rest of the categories. After checking with the dataset, we discovered that that all Bungalows houses are built with Brick and Wood materials and vice versa. This is important to note because including both variables in our regression would suffer from collinearity, as we would encounter later as we run our model.

2.2.4 Construction Time

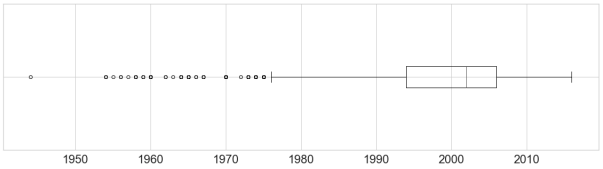
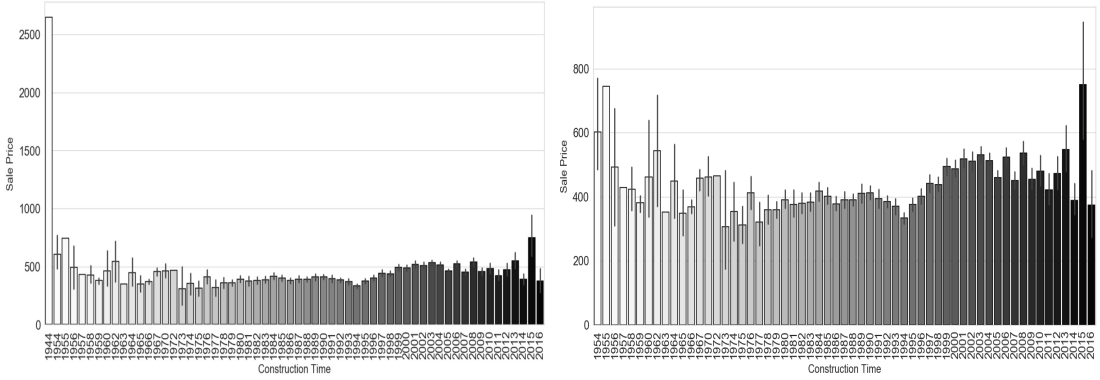


Figure 10: Construction Time and Price



(a) Before (b) After

Figure 11: Construction Time and Price before and after dropping outlier

From the boxplot, we see that there are multiple outliers in the data. After dropping the most significant outlier, we have a more normalized distribution of house prices across the years.

We noticed a slight U shape pattern in our graph, which means that the houses seem more expensive when it is newer or older. This is contradictory to our common sense. After some literature review, we later decided to use $(age + age^2)$ to run our regression, which is detailed in later sections of this paper.

3 Inferential Analysis

3.1 Model Construction

3.1.1 Functional form based on common sense and variable description

When constructing the model for per unit housing prices in Beijing, we apply the "from general to specific" approach. In the first iteration of our model, we introduce all the possible variables as the independent variables. In terms of functional forms, we use $log(price)$ as the dependent variable, and $log(square)$ as the independent variable, for the reasons mentioned above.

Furthermore, as we discussed, we transform the latitude and longitude to the distance between a house and city center, and apply a quadratic term to better represent the bell shape relation between price and distance. Then, we transform all the categorical variables into binary dummy variables. In particular, we use the first category of each categorical variable as the base category, and include the rest of the dummies in our regression.

3.1.2 Functional form based on literature review

"Age has a significantly negative effect while age squared has a significantly positive effect," according to Li and Brown, "houses need to be truly historic before the benefits from their age outweigh the loss in value associated with being older". Therefore, we believe $log(price)$ should have a none-linear relation with construction time (age), i.e. $(A * age + B * age^2)$, which resembles an inverse bell shape, where A is negative and B is positive.

Furthermore, we found that $log(price)$ also should have a non-linear relation with the number of rooms (n), i.e. $(A * n + B * n^2)$, with A being positive and B being negative (Kain and Quigley, 1970). The intuition is that a moderate number of rooms is superior to either too few (inconvenient) or too many (high maintenance cost) rooms.

3.2 First Iteration Regression Model

Source	SS	df	MS	Number of obs	=	10,330
				F(27, 10302)	=	306.05
Model	824.017038	27	30.5191496	Prob > F	=	0.0000
Residual	1027.31887	10,302	.099720333	R-squared	=	0.4451
				Adj R-squared	=	0.4436
Total	1851.33591	10,329	.179236703	Root MSE	=	.31579

price_log	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Lng_diff_quad	-7.801252	.1845434	-42.27	0.000	-8.162993	-7.439511
Lat_diff_quad	-5.981557	.1581843	-37.81	0.000	-6.291629	-5.671485
square_log	-.3232739	.0155626	-20.77	0.000	-.3537795	-.2927683
livingRoom	.1331211	.0148177	8.98	0.000	.1040755	.1621667
livingRoom2	-.0173575	.002662	-6.52	0.000	-.0225755	-.0121394
drawingRoom	.1285189	.0174101	7.38	0.000	.0943917	.162646
drawingRoom2	-.0236394	.0061725	-3.83	0.000	-.0357388	-.01154
kitchen	.098927	.0550441	1.80	0.072	-.00897	.2068241
kitchen2	.0060342	.0300601	0.20	0.841	-.0528895	.0649579
bathRoom	-.0552431	.0238766	-2.31	0.021	-.1020459	-.0084403
bathRoom2	.030211	.0060237	5.02	0.000	.0184033	.0420187
constructionTime_diff	.0109332	.0013903	7.86	0.000	.0082079	.0136585
constructionTime_diff2	-.0000485	.0000298	-1.63	0.103	-.0001069	9.84e-06
ladderRatio	.1684176	.0193751	8.69	0.000	.1304388	.2063965
elevator	.1502676	.0129221	11.63	0.000	.1249377	.1755975
fiveYearsProperty	-.0193993	.0067672	-2.87	0.004	-.0326643	-.0061344
subway	.0927824	.0070136	13.23	0.000	.0790345	.1065304
buildingType2	.6230491	.2747338	2.27	0.023	.0845173	1.161581
buildingType3	.0730704	.0096089	7.60	0.000	.0542351	.0919056
buildingType4	.0971986	.010536	9.23	0.000	.076546	.1178511
renovationCondition2	.0377757	.0215294	1.75	0.079	-.004426	.0799774
renovationCondition3	.0234049	.0131818	1.78	0.076	-.002434	.0492438
renovationCondition4	.104152	.0126615	8.23	0.000	.079333	.128971
buildingStructure2	.0493564	.2239212	0.22	0.826	-.3895726	.4882855
buildingStructure3	0	(omitted)				
buildingStructure4	.0356348	.2243508	0.16	0.874	-.4041363	.4754059
buildingStructure5	.03659	.2423199	0.15	0.880	-.4384041	.511584
buildingStructure6	.1070451	.2240259	0.48	0.633	-.3320892	.5461795
_cons	11.41771	.2329978	49.00	0.000	10.96099	11.87443

Figure 12: First iteration regression model

3.2.1 Test for overall significance

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = 0$$

$$H_1 : \exists \beta_i \neq 0 \text{ for } i = 1, 2, 3...$$

Reject H_0 if $F \geq F_\alpha$, where F_α is the $100(1-\alpha)$ percentile of a $F_{k,n-k-1}$ distribution. Equivalently, we can reject H_0 if the overall *p-value* ≤ 0.05.

$$F_{27,10302} = 306.05$$

$$p\text{-}value = 0.00$$

Therefore, we reject H_0 at 5% significance level, the independent variables are jointly significant. Furthermore, the R^2 of the regression is 0.4451, which indicates that 44.51% of the variance in $log(price)$ can be predicted from the independent variables.

3.2.2 Test for individual significance of β_i

$$H_0 : \beta_i == 0$$

$$H_1 : beta_i \neq 0 \text{ for } i = 1, 2, 3...$$

Given significance level $\alpha = 0.05$, we reject H_0 if *p-value* ≤ 0.05. With the *p-values* given in Figure 12, we cannot reject H_0 for independent variables *kitchen, kitchen2, constructionTime_diff2, renovationCondition2, renovationCondition3, buildingStructure2-6.* Noticeably, *Stata* omitted all test statistics of *buildingStructure3* citing collinearity. Upon careful inspection, we discover that in the original data, the values corresponding to *buildingStructure3* and *buildingType2* are coincidentally identical. Therefore, we decide to also exclude *buildingStructure3* from the model.

3.3 Regression Model Refined

Source	SS	df	MS	Number of obs	=	10,330
				F(18, 10311)	=	454.82
Model	819.364779	18	45.5202655	Prob > F	=	0.0000
Residual	1031.97113	10,311	.100084485	R-squared	=	0.4426
				Adj R-squared	=	0.4416
Total	1851.33591	10,329	.179236703	Root MSE	=	.31636

price_log	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Lng_diff_quad	-7.768248	.1839029	-42.24	0.000	-8.128733	-7.407763
Lat_diff_quad	-6.131044	.1561888	-39.25	0.000	-6.437204	-5.824883
square_log	-.3235396	.015573	-20.78	0.000	-.3540658	-.2930135
livingRoom	.1338677	.0147842	9.05	0.000	.1048878	.1628476
livingRoom2	-.0174411	.0026519	-6.58	0.000	-.0226392	-.0122429
drawingRoom	.138355	.0170525	8.11	0.000	.1049288	.1717812
drawingRoom2	-.0258867	.0060764	-4.26	0.000	-.0377976	-.0139757
bathRoom	-.0487354	.0236405	-2.06	0.039	-.0950754	-.0023953
bathRoom2	.0293854	.0059681	4.92	0.000	.0176868	.0410839
constructionTime_diff	.0084452	.0005022	16.82	0.000	.0074608	.0094295
ladderRatio	.1663154	.0193573	8.59	0.000	.1283714	.2042595
elevator	.1850089	.0102149	18.11	0.000	.1649856	.2050322
fiveYearsProperty	-.0172248	.006644	-2.59	0.010	-.0302482	-.0042013
subway	.0938314	.0070199	13.37	0.000	.080071	.1075918
buildingType2	.5083676	.1592971	3.19	0.001	.1961144	.8206208
buildingType3	.071238	.009554	7.46	0.000	.0525103	.0899657
buildingType4	.0892781	.0102589	8.70	0.000	.0691687	.1093876
renovationCondition4	.0843148	.0064586	13.05	0.000	.0716546	.0969749
_cons	11.62013	.0580308	200.24	0.000	11.50638	11.73388

Figure 13: Refined regression model

3.3.1 Test for overall significance

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = 0$$

$$H_1 : \exists \beta_i \neq 0 \text{ for } i = 1, 2, 3...$$

Reject H_0 if $F \geq F_\alpha$, where F_α is the $100(1-\alpha)$ percentile of a $F_{k,n-k-1}$ distribution. Equivalently, we can reject H_0 if the overall $p\text{-}value \leq 0.05$.

$$F_{18,10311} = 454.82$$

$$p\text{-}value = 0.00$$

Therefore, we reject H_0 at 5% significance level, the independent variables are jointly significant. Furthermore, the R^2 of the regression is 0.4426, which indicates that 44.26% of the variance in $log(price)$ can be predicted from the independent variables.

3.3.2 Test for individual significance of β_i

$$H_0 : \beta_i == 0$$

$$H_1 : beta_i \neq 0 \text{ for } i = 1, 2, 3...$$

Given significance level $\alpha = 0.05$, we reject H_0 if $p\text{-}value \leq 0.05$. With the $p\text{-}values$ given in Figure 13, we can reject H_0 for all independent variables. Therefore, our model is refined.

3.4 Model Diagnostic Tests

3.4.1 Zero conditional mean

Variable	Obs	Mean	Std. Dev.	Min	Max
uhat	10,330	1.38e-10	.315809	-1.391468	1.113695

Figure 14: Summary statistics of $\hat{u}$

The summary statistics generated by *Stata* shows the mean of the model's residual values $\hat{u}$ has a negligible difference from 0, at approximately $1.38e-10$, indicating that our refined regression model has the zero conditional mean property.

3.4.2 $\hat{u}$ is normally distributed

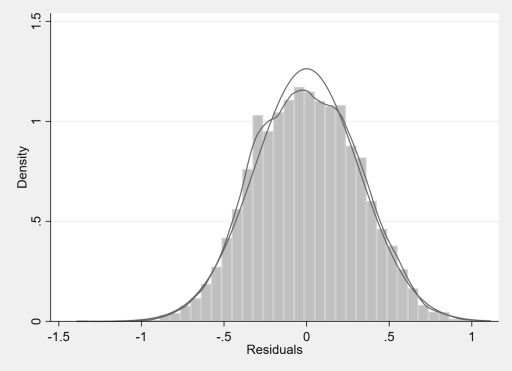


Figure 15: Histogram of $\hat{u}$

The histogram in Figure 14 shows the distribution of fitted residual values $\hat{u}$. The two lines are the fitting curves of normal distribution and kernel probability density function. The distribution of the standardized residuals is approximately identical to a normal distribution.

3.4.3 Fitted values v.s. observed values

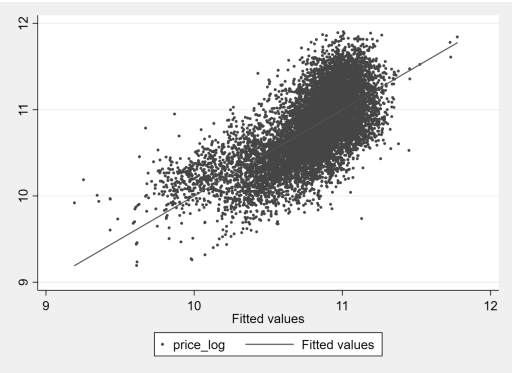


Figure 16: Fitted values v.s. observed values

In Figure 16, we construct a scatter plot to observe the correlation between the observed values of $log(price)$ and the fitted values of $log(price)$. If our model can correctly predict the value of $log(price)$, we should expect to see a 45 degree pattern in the scatter plot, and that is indeed what we observe in the data, where Y-axis is the observed data and X-axis the predicted data.

3.4.4 Residual plot displays heteroskedasticity properties

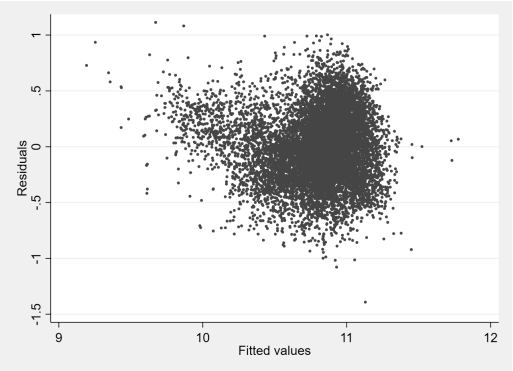


Figure 17: Residual plot $\hat{u}$ v.s. $\hat{y}$

The residuals $\hat{u}$ v.s. fitted values $\hat{y}$ scatter plot displays properties of heteroskedasticity. Under the homoskedasticity assumption, we would expect to see the residuals somewhat evenly or randomly distributed along the fitted values axis. However, in our residual plot, we see the range of residual values noticeably expands for $log(\hat{price})$ close to 11, which means the variance in $\hat{u}$, conditional on the explanatory variables, is not the same for all combinations of the explanatory variables. Under violation of the homoskedasticity assumption, although the OLS estimation remains unbiased, the standard error values are indeed affected and it invalidates the statistical tests of significance which assumes homoskedasticity. Therefore, we first conduct the Breusch-Pagan test for heteroskedasticity to confirm the existence of the issue.

3.4.5 Breusch-Pagan test for heteroskedasticity

Source	SS	df	MS	Number of obs	=	10,330
				F(18, 10311)	=	18.57
Model	5.64821749	18	.31378986	Prob > F	=	0.0000
Residual	174.198896	10,311	.016894472	R-squared	=	0.0314
				Adj R-squared	=	0.0297
Total	179.847114	10,329	.017411861	Root MSE	=	.12998

uhat2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Lng_diff_quad	-.4908019	.0755575	-6.50	0.000	-.6389092	-.3426946
Lat_diff_quad	.024121	.064171	0.38	0.707	-.1016666	.1499087
square_log	-.0315782	.0063983	-4.94	0.000	-.04412	-.0190363
livingRoom	.0092832	.0060742	1.53	0.126	-.0026233	.0211898
livingRoom2	-.0006907	.0010895	-0.63	0.526	-.0028263	.001445
drawingRoom	-.0136014	.0070061	-1.94	0.052	-.0273347	.0001319
drawingRoom2	.0056361	.0024965	2.26	0.024	.0007424	.0105298
bathRoom	-.0338162	.0097128	-3.48	0.001	-.0528552	-.0147771
bathRoom2	.0161978	.002452	6.61	0.000	.0113914	.0210042
constructionTime_diff	-.0013111	.0002063	-6.35	0.000	-.0017155	-.0009066
ladderRatio	-.0013404	.007953	-0.17	0.866	-.0169299	.0142491
elevator	-.0257976	.0041969	-6.15	0.000	-.0340242	-.0175709
fiveYearsProperty	-.0156391	.0027297	-5.73	0.000	-.0209899	-.0102884
subway	.0070068	.0028842	2.43	0.015	.0013533	.0126604
buildingType2	-.1572712	.0654481	-2.40	0.016	-.2855621	-.0289803
buildingType3	-.0051371	.0039253	-1.31	0.191	-.0128315	.0025573
buildingType4	-.0049585	.0042149	-1.18	0.239	-.0132206	.0033036
renovationCondition4	-.0079897	.0026536	-3.01	0.003	-.0131912	-.0027882
_cons	.2971915	.0238423	12.46	0.000	.250456	.343927

Figure 18: Breusch-Pagan test for heteroskedasticity

To conduct the Breusch-Pagan test for heteroskedasticity, we predict the residuals from our refined model regression, and regress all explanatory variables on the square of residuals.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = 0$$

$$H_1 : \exists \beta_i \neq 0 \text{ for } i = 1, 2, 3...$$

We can reject H_0 if $F \geq F_\alpha$, where F_α is the $100(1-\alpha)$ percentile of a $F_{k,n-k-1}$ distribution. The *F-statistic* of the model is just the *F-statistic* for the Breusch-Pagan test.

$$F_{18,10311} = 14.25$$

Therefore, we reject H_0 at 5% significance level, the independent variables are jointly significant on the square of residuals, which means our model fails the B-P test and has heteroskedasticity issue. However, the R^2 of the regression is 0.0316, which indicates that only 3.16% of the variance in $\hat{u}^2$ can be predicted from the independent variables. In order to eliminate the influence of heteroskedasticity on standard errors and significance tests, we apply the heteroskedasticity-robust standard errors in the next iteration of OLS estimation.

3.5 Heteroskedasticity Robust Version

Linear regression

Number of obs	=	10,330	
F(18, 10311)	=	471.59	
Prob > F	=	0.0000	
R-squared	=	0.4436	
Root MSE	=	.31608	

price_log	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Lng_diff_quad	-7.772003	.1794032	-43.32	0.000	-8.123668	-7.420338
Lat_diff_quad	-6.045975	.1498406	-40.35	0.000	-6.339692	-5.752259
square_log	-.3266811	.0171551	-19.04	0.000	-.3603085	-.2930537
livingRoom	.1266209	.0151061	8.38	0.000	.0970101	.1562318
livingRoom2	-.0160333	.0028228	-5.68	0.000	-.0215666	-.0105001
drawingRoom	.1380134	.0178715	7.72	0.000	.1029818	.1730451
drawingRoom2	-.0269155	.0065875	-4.09	0.000	-.0398282	-.0140028
bathRoom2	.0177667	.0033566	5.29	0.000	.0111872	.0243462
constructionTime_diff	.0090475	.0005247	17.24	0.000	.0080189	.0100761
ladderRatio	.1711196	.0219524	7.80	0.000	.1280887	.2141505
elevator	.1469463	.0148498	9.90	0.000	.1178378	.1760549
fiveYearsProperty	-.016668	.006749	-2.47	0.014	-.0298974	-.0034386
subway	.0928612	.0069073	13.44	0.000	.0793215	.1064008
buildingType2	.5371018	.0414938	12.94	0.000	.4557659	.6184377
buildingType3	.0712969	.009512	7.50	0.000	.0526515	.0899423
buildingType4	.0931076	.010597	8.79	0.000	.0723355	.1138797
renovationCondition4	.0832504	.0064659	12.88	0.000	.0705759	.0959248
buildingStructure6	.0563847	.0138637	4.07	0.000	.0292091	.0835602
_cons	11.57876	.0659371	175.60	0.000	11.44952	11.70801

Figure 19: Heteroskedasticity robust regression model

Using the heteroskedasticity-robust standard errors in OLS estimation, we can conduct similar significance tests as before. Independent variables *kitchen, kitchen2, bathRoom, constructionTime_diff2, renovationCondition2, renovationCondition3, buildingStructure2-5* are found to be not statistically significant to $log(price)$ and are excluded. Finally, we arrive at the model:

$$\begin{aligned}\hat{price\_log} = &-7.77 * Lng\_diff\_quad - 6.05 * Lat\_diff\_quad - 0.33 * square\_log \\ &+ 0.13 * livingRoom - 0.02 * livingRoom^2 + 0.14 * drawingRoom \\ &- 0.03 * drawingRoom^2 + 0.02 * bathRoom^2 \\ &+ 0.01 * constructionTime\_diff + 0.17 * ladderRatio + 0.15 * elevator \\ &- 0.02 * fiveYearsProperty + 0.09 * subway + 0.54 * buildingType2 \\ &+ 0.07 * buildingType3 + 0.09 * buildingType4 \\ &+ 0.08 * renovationCondition4 + 0.06 * buildingStructure6 + 11.58 \end{aligned} \tag{1.0}$$

4 Conclusion

4.1 Interpretation of Our Model

This paper constructs a multi-variable linear regression model to answer the question "what are the key factors that influence per unit housing prices in Beijing?". For proper elasticity interpretation, we set $log(price)$ as the dependent variable.

First, we exclude variables that are related to the website and obviously irrelevant to the housing prices, such as id and link to the data. Then, we compute the descriptive statistics of the remaining variables, based on which we made certain adjustment to the functional form of certain variables for interpretation correctness.

Following the variable description, we introduce the independent variables. The chosen variables include 9 numerical variables, 3 dummy variables, and 3 categorical variables which we then transformed into multiple dummy variables.

Next, we attempt to refine our model's functional form by making statistical analysis of the variable description and correlation matrix, as well as consulting previous research papers on the commonly adopted approach when it comes to constructing house price predicting models.

We introduce $bedroom^2$, $livingroom^2$, $drawingroom^2$, $bathroom^2$, $kitchen^2$ to our model, based on our literature review, to better model the effect of room numbers on per unit price. What's more, previous research also reveals an interesting relationship between the house's age and its price, which resembles a U-shape curve. It contradicts to our common sense by indicating higher price for older construction. Thus, we also introduce age^2 into our model to further explore the significance of this pattern.

In the inferential analysis section, we first construct Model I based on our previous conclusions regarding variable choice and functional form. By testing the model for overall significance and individual variable significance, we are able to verify our assumptions and exclude statistically insignificant variables from our model. *kitchen, kitchen2, constructionTime_diff2, renovationCondition2, renovationCondition3, buildingStructure2-6* are drop after this step.

Our refined Model II shows statistical significance of all independent variables, but failed to pass the Breusch-Pagan test of heteroskedasticity, which is supported by the residual plot (residuals v.s. fitted values). Since the heteroskedasticity is of unknown form, we decide to use the heteroskedasticity-robust form of OLS regression to re-test our original model. After a similar significance test and consequent exclusion of variables, we reach our final model as shown in Formula (1.0). The final model has $R^2 = 0.4436$ and $F(robust) = 471.59$, which is valid in

our large samples. Moreover, various model diagnostic tests return very similar results as in the non-robust scenario.

Now we are able to answer our research question based on our model: **what are the key factors that influence per unit housing prices in Beijing?**.

- Longitude and latitude difference from the city center (i.e. *distance*) has a huge negative effect on the per unit housing price. The coefficient between $log(price)$ and $longitude\ distance^2$ (-7.77) is by far one of the largest in absolute value. Coupled with the quadratic form the longitude distance variable, we can see housing price drops very fast as its location moves away from the city center. Similarly, the coefficient between $log(price)$ and $latitude\ distance^2$ (-6.05) also suggest a strong relationship (though weaker than longitude) between latitude distance and price. Combining these two findings, one can suggest that the location of the house is a dominant factor affecting the housing price in Beijing.

- Dummy variable $buildingType2$ shows considerable impact on $log(price)$. Holding all other variables constant, the coefficient (0.54) indicates that price will increase by $(e^{0.54} - 1) = 71.6\%$ if the house is a bungalow, which typically represents Siheyuan, relative to a tower.

- Independent variable $log(square)$ has a coefficient of -0.33, which suggests that for every 1% increase in the total square meter of the house, it leads to a 0.33% decrease in price.

- The number of drawing-rooms is found to have a non-linear effect on price. In the range of [0.5] drawing rooms, we observe the function $+0.14 * drawingroom - 0.03 * drawingroom^2$ in our regression, which indicates that, at a small number of rooms, an additional drawing room has a positive effect on $log(price)$. However, at FOC $= 0 \rightarrow x = 2.33$, the effect becomes negative, and the bell-shaped curve means that the optimal number of drawing rooms is around 2.

- Similarly, the positive coefficient of *living room* and the negative coefficient of $living\ room^2$ show that the number of living room does have a significant effect on price, but as the number of living rooms increase, its effect goes from positive to negative. In the range of [0.9] living rooms, FOC $= 0 \rightarrow x = 3.25$ suggests that the optimal configuration is having around 3 living rooms.

- The coefficient of $bathroom^2$ (0.02) shows an increasing contribution of each additional bathroom to the per unit house price.

- To our surprise, the age of construction only has a weak positive effect on price. The coefficient shows that the unit house price increase by 1% for every 1 year increase in age.

- As expected, as the ladder ratio increase by 1, price increase by 18.5% on average; availability of elevator increases price significantly by 16.2%; while availability of subway adds value to the house by 9.4%.

4.2 Critical Thoughts

We notice that some parts of our modeling result is contradictory to the hypothesis we proposed (see section 2.2.4), which suggests that $log(price)$ should have a non-linear relation with $(A * age + B * age^2)$, where A is positive and B is negative (Kain and Quigley, 1970). In our model, variable age^2 ($constructionTime\_diff^2$) is excluded since it failed to pass the significance test at 5%, and it is in fact the least significant variable in our model.

This finding suggests that the effect of age on house price in Beijing is not only monotone, but positive as well. To be specific, the house price increase by 1% for every 1 year increase in age, unlike the prediction made by Li and Brown that price benefits from either a small (new) or large (historic) age factor. We make the observation that older houses tend to be located in

the central areas of a city, contributing to their high prices. Since Beijing is a concentric city, old houses located in the city center can offer a lot more convenience to people as the trend of urbanization moves outward, and people value the convenience more than the physical condition of the house (i.e. when it was built). This inconsistency is also comprehensible if one think of Siheyuan and Hutong in Beijing, which are of high historical value and are mainly located in the city center, subsequently being perceived as precious and expensive. Although the number of observations of such cases is small in our dataset, they could have considerable impact on the coefficients and the intercept of our model.

4.3 Limitations and Further Work

An important limitation of our current study is the observed heteroskedasticity property, which is most likely due to omitted variables in our final model. However, since we started using all available variables as the independent variables, and the fact that many possible quadratic terms and log terms are being included, we can be confident that no variables in our dataset is mistakenly excluded. Therefore, it is very likely that the data provided in the dataset, that is collected from the Lianjia website, does not include all the factors that may affect price.

4.3.1 Methodology limitation and suggestion

The heteroskedasticity issue is one we need to fully acknowledge. Under heteroskedasticity, OLS is no longer the best linear unbiased estimator (BLUE), despite being unbiased and consistent. There may be more efficient linear estimators available, such as the weighted least square (WLS) estimator and the feasible generalized least squares (FGLS) estimator. Nonetheless, since the heteroskedasticity in this case is not known up to a multiplicative constant, and that we have no information about the possible form of the heteroskedasticity function, it could be demanding to apply these methods, which is left to be done in the future.

4.3.2 Information limitation and suggestion

Econometric research is almost always backed up by solid economic theories and common sense. Therefore, careful derivation of the economic model of price can take us one step closer to finding the true model that determines housing price in Beijing. Upon careful research, we learned that most researchers build their theoretical model for housing price by using the hedonic pricing principle as reference (Rosen, 1947), under which the unit price of a house is determined by 3 categories of attributes: structure, location, and neighborhood (Yusof and Ismail, 2012). Thus, we believe that one significant deficiency of our model stems from the lack of sufficient information regarding some aspects of the house. For future improvements, more data is needed.

Housing structure: Houses that sit on a higher floor tend to be more expensive compared to those on the first or the second floor. Furthermore, Feng Shui is sometimes an important factor in Chinese people's decision-making, and south-facing houses are typically more valued.

Housing location: Distance to the city center may have limited explanatory power if it is the only measure of location. Distance to CBD and sub-centers can be a better measure of good location (Chen and Hao, 2006). Moreover, the school district that the house is in is sometimes a highly sought-after characteristic, which is especially true in big cities like Beijing.

Neighborhood: Income and housing price are generally highly correlated (Woodridge, 2012). Information such as school quality, crime rate, and residential density can sometimes play an essential role in determining the price.

To conclude, more extensive data are required to estimate both the quantitative and qualitative attributes of the houses. Meanwhile, an increment to the sample observations could add appreciably to the sample information, and enable better estimation of the price model.

5 Modeling Environment

- Stata SE
- Python 3.7.6

6 References

Chen, J. and Hao, Q.J. (2006). "Housing Market Development and Housing Affordability in Shanghai 1993-2005". Paper for Uppsala-Tsinghua Joint Conference on "Housing Affordability in China", Beijing, April.22-24, 2006.

Fletcher,M., Gallimore,P. Mangan,J., (2000). Heteroscedasticity in Hedonic Price Models. Journal of Property Research, 17(2), pp. 93-108.

John F. Kain John M. Quigley (1970) Measuring the Value of Housing Quality, Journal of the American Statistical Association, 65:330, pp. 532-548.

Nicodemo, Catia Raya, Josep Maria, 2012. "Change in the distribution of house prices across Spanish cities," Regional Science and Urban Economics, Elsevier, vol. 42(4), pp. 739-748.

Yusof, A. and Syuhaida Ismail. "Multiple regressions in analysing house price variations." Communications of The IbIMA (2012): 1-9.